Cloud Services Performance Management

Based on Network Traffic Prediction

Alexandra Varfolomeeva*[1], Victor Romanov[2]

Department of Information Systems in Economics and Management, Russian Plekhanov University of Economics Moscow, Russian Federation

*[1]aovarfolomeeva@gmail.com; [2]victorromanov1@gmail.com

Abstract

Cloud computing is a rather new and extremely popular concept in IT industry and there are still a significant number of challenges that need to be addressed. The main goal of cloud computing provider is to maximize its profit by minimizing costs and the amount of violations to Quality-of-Service (QoS) levels agreed with customers.As the network is considered to be a backbone and bottleneck of cloud services, it is important to predict the number of service requests and to manage the performance of cloud applicationsacross the diverse cloud deployment models. Unfortunately, it is very difficult to predict the demands on the network and accurate traffic models are necessary for service providers to properly maintain quality of service and to distribute workload across one or more servers, network interfaces, hard drives, or other computing resources. In this paper,it isproposedthat branching processes theory will appropriatefor description ofthe dynamics of cloud services demand as well as prediction of the network traffic in cloud computing environment. Branching stochastic processes are generally used to describe random systems such as nuclear chain reactions, population development, epidemic of disease spread and gene propagation. It isshowed through simulation and experiment that cloud computing demand can be developed as a branching stochastic process and our approach to cloud services performance management is able to dynamically adapt to time-varying workloads and to reduce QoS violations.

Keywords

Cloud Computing; Branching Process; Network Traffic Modeling; Performance Management

Introduction

Cloud computing is a new model of delivering computing resources in which centrally administered computing capabilities are provided as services on-demand over the network to a variety of customers. According to IDC's analysis, it is forecasted that the worldwide cloud services in 2013 will amount to $44.2bn, of which the European market will reach €6,005m in 2013.[1]As the network is the backbone and the bottleneck of cloud computing, the prediction onthe number of service requests coming to appear and managing the huge number of operations and volumes of data within a cloud transparently and without service interruptions is a criticaltask.In this paper,it isproposed that cloud services performance management model would be helpful for cloud service providers to represent the dynamics of cloud services demand by stochastic processes, and to distribute workloads based on prediction across different types of computing resources, in order to improve the quality of service provided.It issuggested that branching processes theory will appropriatefor both description onthe dynamics of cloud services demand and prediction onthe network traffic in cloud computing environment. The rest of this paper is organized as follows: a literature review is available in Section 2; Section 3 depicts the basic principles of cloud computing; introduction ofbranching processes theory will be shown in Section 4; Section 5 provides cloud services performance management model;Section 6 presents the simulation experiment results and comparative analysis; the conclusion will be made in the last section.

Literature Review

The design of robust and reliable network services for cloud computing environment is a challenging task. The only path to achieve this goal is to develop a deepunderstanding onthe traffic characteristics. An accurate estimation of the network performance is vital. According to Ahmed M. Mohammed and Adel F. Agamy,traffic models enable network designers to make assumptions onthe networks based on previousexperience and prediction of performance for future rapidly changing requirements in cloud environment. A corpus of literature on network traffic modeling exists. One of the most widely used and oldest traffic models is the Poisson Model is characterized as a renewal process and Poisson distribution is the predominant model used analyze traffic in traditional telephony networks. Deterministic

[1]Frank Gens, Robert P Mahowald, Richard L Villars. "IDC Cloud Computing 2010 - An IDC Update."Doc # TB20090929, 2009

Traffic Model is proposed to providereal time service over real time channel where clients declare their traffic characteristics and performance requirements at the time of channel establishment in this model. Chaotic maps are low dimensional nonlinear systems whose time evolution is described based onknowledge of an initial state and a set of dynamical laws. In "Chaotic Maps as Models of Packet Traffic: The Fundamental Role of Teletraffic in the Evolution of TelecommunicationsNetworks" traffic characteristics modeledby means of consideration of chaotic maps are illustrated. Wavelet-based models use wavelet transform function to model long-range dependence traffic such as traffic measured on Ethernet. Multifractal wavelet model is presented in"Simulation of non-Gausian Long-Rang-Dependent Traffic Using Wavelets".Researchers Hao-peng Chenand Shao chong Liin their paperare going in the same direction with us. In this model, the web applications are presented as queues and the virtual machines are modeled as service centers. The queueing theory is applied to explore the wayhow virtual machines are dynamically created and removedin order to implement scaling up and down.

Cloud Computing

Cloud computing is a model capable of convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly providedand released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. [2]

There are three service models of cloud computing:

Software as a service (SaaS): is software offered by a third party provider, available on demand, usually via the Internet configurable remotely. Examples include online word processing and spreadsheet tools, CRM services (Salesforce CRM, Google Docs).

Platform as a service (PaaS): allows customers to develop new applications using APIs that aredeployed and configurable remotely. The platforms offered include development tools, configuration management, and deployment platforms. Examples are Microsoft Azure and Google App engine.

Infrastructure as service (IaaS): provides virtual machines and other abstracted hardware as well asoperating systems. Examples include Amazon EC2 and S3, Terremark Enterprise Cloud, Rackspace Cloud, and Onlanta.

The following deployment models are available for cloud computing services:

- Private cloud: services built according to cloud computing principles, but accessible only within a private network
- Community cloud: cloud services offered by a provider to a limited and well-defined number of parties
- Public cloud: public available - any organization may subscribe
- Hybrid cloud: a composition of two or more clouds (private, community or public)

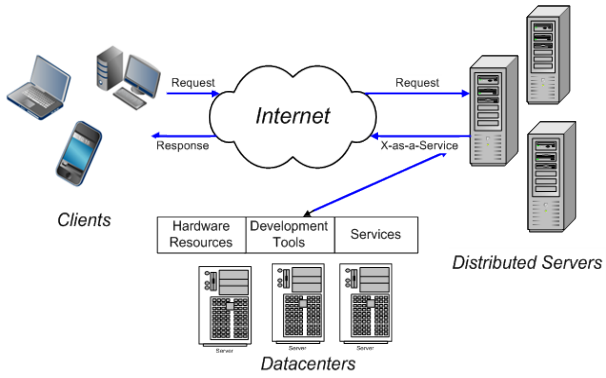

FIG. 1 CLOUD COMPUTING TOPOLOGY

Thus, cloud computing provides a pool of highly scalable and easily accessible virtualized resources capable of hosting end-user applications exploited in a pay-as-you-go model. According toTsaravas C. and Themistocleous M., cloud computing involves the following three basic components which are illustrated in FIG. 1: clients, datacenter and distributed servers. For many companies with highly variable IT needs, cloud computing can be an alternative to an expensive oversupply of in-house computing power. However, there are some major obstacles which hinder the adoption and growth of cloud computing. As alltechnological concept, cloud computing is not an exception in terms of trust and security issues. Once data are outsourced to a third-party cloud provider, several concerns arise about security, availability and reliability of data.

Branching Processes Theory

A branching process is a process where an initial random number of objects 'create' more objects of the same or different type, and these objects continue to

[2] NIST: Cloud Computing Program:http://www.nist.gov/itl/cloud/index.cfms.Accessed Nov 2012.

'create' other objects, with the system developedin accordance with some probability law. Branching processes are used to describe random systems such as population development, nuclear chain reactions and spread of epidemic disease. An example of such a process is a population of individuals developing from a single progenitor – the initial individual.andproducinga random number of offspring, each of whichin turn produces a random number of offspring; and so the process continues as long as there are alive individuals in the population. FIG. 2 is a graphic illustration of a general multilevel branching processwhich was proposed by Galton F.,in addition,the probability of extinction was first obtained by Watson H., taking into account the probability generating function for the number of children in the nth generation.

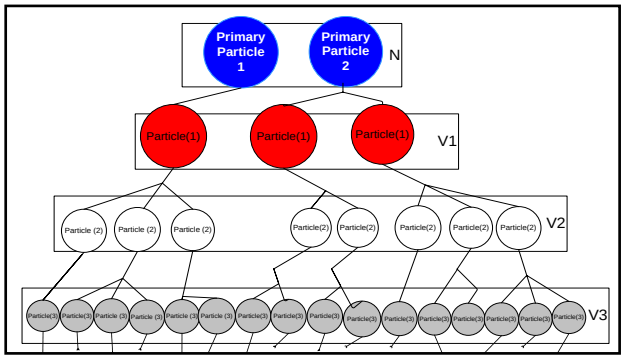

FIG. 2 GRAPHIC ILLUSTRATION OF A MULTILEVEL BRANCHING PROCESS

Let *N*be a random variable, with a probability distribution function $g_N(n) = P(N = n)$, with mean μ_N and variance σ_N^2 . Let *{Xi}*be a series of independent identically distributed random variables, with a common distribution$f_X(X)$and with μ_X as the mean andσ_X^2 as the variance of each element in the series. The sum of *N* elements of the series $\{X_i\}$ is denoted by the following sum

$$V = X_1 + X_2 + ... + X_N \quad (1)$$

The mean of V is denoted by

$$E[V] = E[X_i] * E[N] \quad (2)$$

And the variance of V is denoted by

$$\mathrm{Var[V]} = \mathrm{E[N]} * \mathrm{Var[X_i]} + \mathrm{Var[N]} * \mathrm{E^2[X_i]} \quad (3)$$

The distribution density function $f_V(v)$ of V can be derived from the basic formula for conditional probabilities

$$f_V(v=k) = P(V=k) = \sum_{n=0}^{\infty} P(N=n)P(X_1+...+X_n = k) \quad (4)$$

Let us denote $f_X(X)$ and $\varphi_X(\omega)$ as the distribution density function and generation function of X_i respectively, and $g_N(n) = P(N = n)$, $\varphi_N(\omega)$ as the distribution density function and generation function of *N*, respectively. For a fixed *n*, the distribution of the sum $X_1 + ... + X_N$ is expressed by the *n*-fold convolution of $\{ f_X(X) \}$ with itself, due to the independence of the series $\{X_i\}$. Equation (4) can be written in a more compact form

$$f_V(v = k) = \sum_{n=0}^{\infty} g_N(n) \{f_X(X)\}^n \quad (5)$$

This formula can be simplified by using the generating functions.

Branching processes theory can be applied tomodeling in cloud computing environment. One of the most important issues in cloud computing environment concerns network efficiency and performance prediction. Where there are large quantities of data involved in an application, access to the data must be fast and reliable or the application's runtime will be excessive. From the viewpoint of a service provider, demands on the network are not entirely predictable. Branching processes theory helps us to model the dynamic demands for cloud services. More and more clients are informed ofthe service, one client from another based on random mechanism. This process is similar to epidemic of disease spread.

Cloud Services Performance Management Model

One of the important challenges cloud service providers confront is the effective management of cloud services performance. The ultimate goal of a cloud service provider is the maximization of its profit through reducing the number of QoS violations and decreasingservice costs. As illustrated in FIG. 3, on the one hand, resource over-provisioning helps to achieve quality of service levels, but itsignificantly increases the service cost. On the other hand,resource under-provisioning helps to reduce costs, but increasingthe possibilities ofQoS violations.

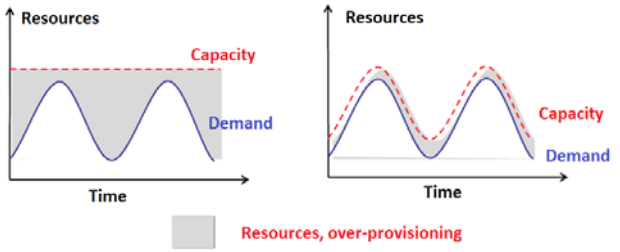

FIG. 3 CLOUD PROVIDER'S RESOURCES MANAGEMENT

In this paper, we present cloud services performance management modelcapableto predict the frequency of

queriescoming to appear based on branching processes theory, in order to maximize the profit of cloud service provider by minimizing QoSviolations. At the same time, this model helps to reduce the resources consumed by physical infrastructure through implementing load balancing solutions. Load balancing (FIG. 4) provides newopportunities for resource managementin the cloud. Amechanism being provided enables the monitoring and measurement of service requests received to ensure that SLAs (Service Level Agreement) are met and clients can obtainvalue from the services they havereceived.

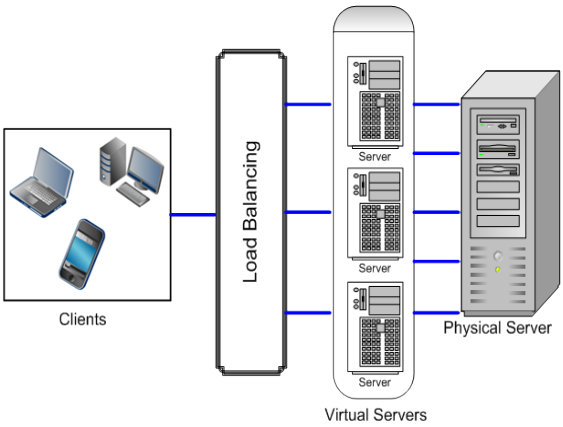

FIG. 4 LOAD BALANCING ARCHITECTURE

Cloud service consumers have a goal to choose the most preferred service they can afford. The utility function for each service user is denoted by

$$f(\lambda_i, t) = -\frac{a_i(t)}{\lambda_i},\ i=1,\ldots,N, \qquad (6)$$

where$N$ is the number of service clients, $\lambda_i \gg 0$is the intensity of user's requests submission.

The utility maximization problem is expressed as

$$f(\lambda_i, t, P) = -\frac{a_i(t)}{\lambda_i} - c_0\lambda_i - c_0\lambda_i P,\ i=1,\ldots,N, \qquad (7)$$

wherec_0 is the network traffic price, $P$the probability of queries rejection,$a_i(t)$user's network traffic needs.

Cloud service providers try to ensure the quality of service provided and to reduce the service response time.

$$T = \sum_{i=1}^{n} \omega_i \tau_i \rightarrow min\,, \qquad (8)$$

where ω_i is the number of queries $\tau_i$is the service time. The cost of processing the query should not exceed v_0:

$$\sum_{i=1}^{n} \omega_i c_i \ll v_0 \qquad (9)$$

FIG. 5 contains a high-level description onthe architecture of cloud services performance management model.

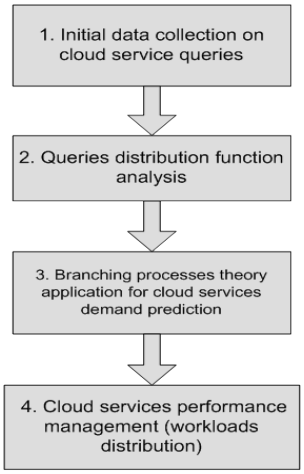

FIG. 5 THE FRAMEWORK OF CLOUD SERVICES PERFORMANCE MANAGEMENT MODEL

To design availableand efficient network solutionsto cloud environment and to understand and solve performance problems arising in communication networks, providers require accurate models to describe network traffic. The main problem is to forecast the frequency of queries further to appear. To evaluate the performance of the proposed technique, the wayis demonstrated how branching processes theory can be applied tobuild data flow prediction models.

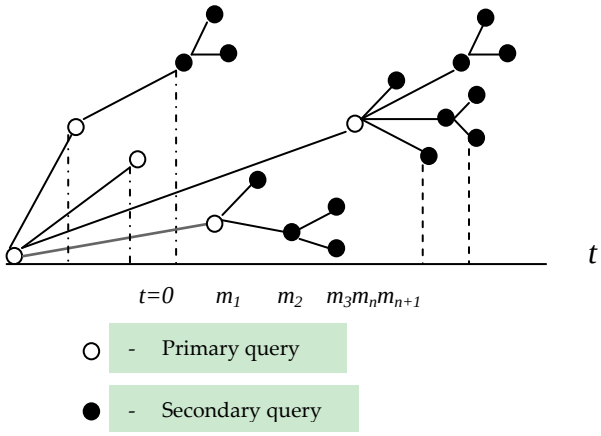

FIG. 6 A BRANCHING PROCESS OF GENERATING QUERIES

The initial vertex of the graph is assigned as *t=0* (FIG. 6), time when the original message reaches the host *Ci*. This point is taken as the reference time of receipt of queries to the cloud service provider's host center*Ci*. The initial vertex of a graph is based on the number of arcs equal to the quantity of primary need. Arcs whose vertices correspond to the secondary queries come of the vertices of the graph corresponding to the initial query. It is considered that the primary queries come into the provider's host randomly. The process of queries admission will be considered as a branching, while allowing that individual queries' paths are

independent. The time intervals between *t=0* and the time of requests admission are random variables with distribution function $F_1(x)$ and density $f_1(x)$. Obviously, $f_1(x)$ represents the latency of the processes of information alerts. The number of initial requests coming to the provider for a certain period of time is the random variable

$$v_1, P(v_1 = k) = p_{1k}, k = 0,1,2 \ldots; \sum_{k=0}^{\infty} p_{1k} = 1 \quad (10)$$

It is assumed that the distribution of the primary query is binomial. The generating function corresponding to the binomial distribution of the primary query is:

$$G_1(u)=(q_0+q_1u)n, \quad (11)$$

where u is the parameter of the generating function. The distribution function of the moments of primary queries receipt is exponential $F1(x)=1-e-\lambda_1 x$, where $\lambda_1$is the volume of primary queries. After each initial query with probability p0 does not appear any secondary query, and with probability $p_1=1-p_{0,}$ there is at least one secondary request. The distribution of secondary queries generated by one primary request is assumed to be subject to the binomial distribution with generating function:

$$G_2(u)=(p_0+pu)n, \quad (12)$$

where*u* is the parameter of the generating function. The distribution function of the time intervals between the moments of initial queries receipt and thatstimulated directly by their secondary queries is defined as $F_2(x)=1-e-\lambda_2 x$, where λ_2 is the intensity of secondary queries receipt. The expectation onthe number of requests in the time interval [t, t+τ] is:

$$M^*[t,\tau]=\begin{cases}\dfrac{nq_1\lambda_1\tau}{\lambda_1-\lambda_2 p_0}\left\{(\lambda_1-\lambda_2)e^{-\lambda_1 t}+\lambda_2 p_1 e^{-\lambda_2 p_0 t}\right\} \\ \qquad\qquad \text{я } (\lambda_1 \neq \lambda_2 p_0) \\ \dfrac{nq_1\lambda_1\tau}{p_0}e^{-\lambda_1 t}(p_0+p_1\lambda_1 t \qquad \text{і } \lambda_1=\lambda_2 p_o\end{cases} \quad (13)$$

If $\frac{\lambda_1}{\lambda_2} < p_1$, then the distribution has a maximum. If$\frac{\lambda_1}{\lambda_2} \geq p_1$, then the distribution hasn't a maximum and decreases from the beginning. In case $\lambda_1 = \lambda_2 p_0$, for $p_0 \geq p_1$ distribution monotonically decreases; for $p_0 < p_1$ distribution has a maximum. Expectation function graphs of the number of queriesis constructed for the following values: time units are *t=[0,130]*, the number of consumers which may create the primary requests is 10, the probability of the customer application with the primary request is $q_1=0.7$, the time interval of the request arrival $\tau=1$, the volume of the primary queries from the consumer at a time $\lambda_1=0.05$. The nature of the functional dependence is affected by primary and secondary queries intensity compliance. FIG. 7 contains plots of the expectation of the number of queries, in this case, the probabilities of the primary p_0 and secondary p_1 queries: for $p_0 \geq p_1$ distribution monotonically decreases, for $p_0 < p_1$ the distribution has a maximum.

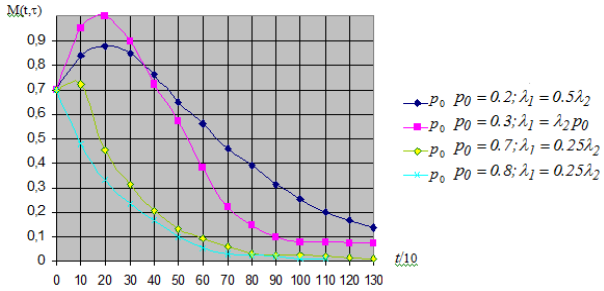

FIG. 7 THE THEORETICAL QUERIES DISTRIBUTION FUNCTIONS OVER TIME

Experiments and Evaluation

Global giants like Amazon.com are definitely more cost-effective and less sensitive to daily, annual and other traffic imbalances than the local providers. In order to evaluate the performance of the proposed technique and to explorethe further need for the model we develop, the Russian SaaS market has beenanalyzed. Megaplan (www.megaplan.ru), Russian Saas provider, encountered difficulties in providing a comprehensive explanation onthe dynamic characteristics of network traffic and had even experienced the bottlenecks that caused provider to lose money. To measure how good the cloud services performance management model is, experiments have beencarried out where cloud usage logs have beenanalyzedto examine whether our model can predict the loads seen in the logs.Megaplan currently offering the following products: Collaboration Package, CRM (Clients and Sales) and Business Manager.Firstly, all traffic data is collected and measuredin order to ensure that cloud services are delivered withinagreed SLA targets. As shown in FIG. 8, all query flows are divided into two separate traffic classes:business service and standard service; in whichbusiness class services generate more revenue. The number of business class requests is about 10%, and their service is more expensive and in case of idle

capacity the losses are unacceptable.

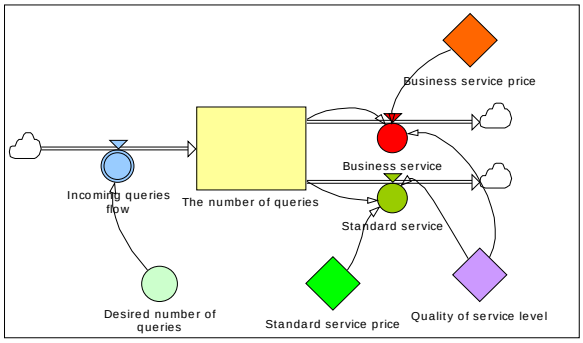

FIG. 8 BUSINESS AND STANDARD SERVICE CLASSES

FIG. 9 demonstrates our configured service item queries flow intensities. It isassumed that all customers are within the same time zone, so that our model represents a reasonably accurate representation of user intensity level throughout a given 24-hour period (X – time, where the graph series (1, 2, 4, 6, 8, 10, 12) match the following time intervals (00:00, 04:00, 08:00, 12:00, 16:00, 20:00, 00:00), Y – incoming queries rate). All the simulations are performed using MATLAB 7.12.0 (R2011A)).

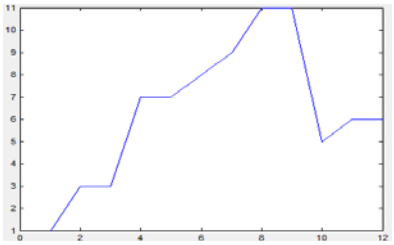

FIG. 9 INCOMING QUERIES INTENSITY

FIG. 10 shows properties and behaviour of branching processes.While FIG. 11 represents queries flow distribution function over time.

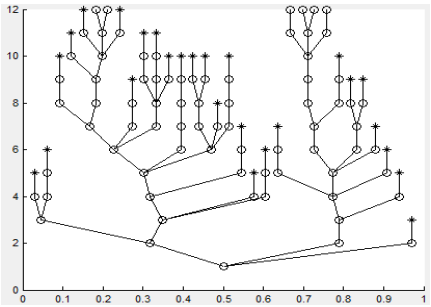

FIG. 10 QUERIES TREE VIEW

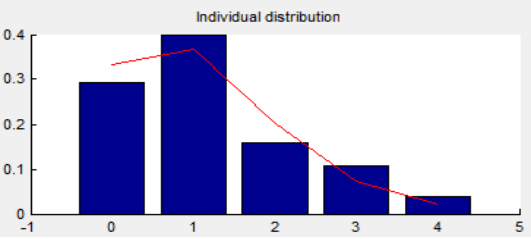

FIG. 11 QUERIES FLOW DISTRIBUTION

Network traffic prediction is the key to managecloud services performance and maintain the quality of service on a high level.Though the researchers Hao-peng Chen and Shao-chong Li in their paper "A queueing-based model for performance management on cloud" are moving in the same direction, in this paper the focus is on the problem not only from the point of view of queueing theory but alsowith relation tomicroeconomics. Unlike the above authors, we do not use service time as a key criteria. Our model is based on profit-driven approach, where profit is the result of the service provided and service time is considered as a constraint.In our model,incoming queue of tasks is a result of branching process that provides us possibility to take into account changein time flow distribution parameters. The economic efficiency of SaaS applications is the focus. The essential part of the model is that it contains parameter makingit possible to adjust the rate of business and standard tasks being served toobtain a higher level of profit. Therefore, to some extent our model can be regarded as system with feedback loop.

Conclusion and Future Work

Cloud computing is a model to deliver computing resources, in which centrally administered computing capabilities are provided as services on-demand over the network to a variety of customers. As popularity of cloud services is growing rapidly, cloud-service providers must guarantee that data are processed effectively and transferred when and where they are needed. Unfortunately, it is extremely difficult to predict the exact performance characteristics and demands on the network at any particular time.

In this paper, cloud services performance management model is put forward that would be helpful for cloud service providers to represent the dynamics of cloud services demand by stochastic processes, and to distribute workloads based on prediction across different types of computing resources, in order to avoid bottlenecks and improve the quality of provided service. It is based on profit-driven approach, where profit is the result of the service provided and service time is considered as a constraint.

REFERENCES

Ahmed M. Mohammed & Adel F. Agamy. "A Survey on the Common Network Traffic Sources Models." International Journal of Computer Networks, Vol. (3): Issue (2), 2011.

Amazon Amazon Elastic Compute Cloud. http://aws. amazon.com/ec2/. Accessed Nov 2012.

Amazon: Amazon Simple Storage Service. http://aws.amazon.com/s3/. Accessed Nov 2012.Cohen I., Golan R., Rotman S. "Applying Branching Processes Theory for Building a Statistical Model for Scanning Electron Microscope Signal." Optical Engineering 39(01), Jan 01, 2000.

Erramilli, Singh R.P., and PruthiP. "Chaotic Maps as Models of Packet Traffic: The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks." Proc. of the 14th ITC, 6-10 June, 1994.

Ferrariand D., Verma D. "A Scheme for Real Time Channel Establishment in Wide Area Networks." IEEE Journal On Selected Areas in Communications, 8(3), April 1990.

Galton F."Problem 4001." Educational Times, April 1, 1873.

Hao-peng Chen Sch. of Software, Shanghai Jiao Tong Univ., Shanghai, China, Shao-chong Li. "A Queueing-Based Model for Performance Management on Cloud." Advanced Information Management and Service (IMS), 2010 6th International Conference, Nov. 30 2010-Dec. 2 2010.

Harris T.E. "The Theory of Branching Processes", Springer, Berlin, 1963.

Megaplan:http://www.megaplan.ru/. Accessed Nov 2012.

Onlanta: http://onlanta.ru/. Accessed Nov 2012.

Rackspace Cloud: http://www.rackspace.com/cloud/. Accessed Mar 2012.

Salesforce CRM: http://www.salesforce.com/eu/. Accessed Mar 2012.

Terremark Enterprise Cloud: http://www.terremark.com/services/infrastructure-cloud-services/enterprise-cloud.aspx. Accessed Nov 2012.

Tsaravas C. and Themistocleous M. "Cloud Computing& E-Government Myth or Reality?" tGov Workshop '11, March 17 – 18, 2011.

Victor Romanov, Aleksandra Varfolomeeva, Andrey Koryakovsky. "Branching Processes Theory Application For Cloud Services Demand Modeling Based on Traffic Prediction." CAiSE 2012 International Workshops, Gdańsk, Poland, June 25-26, 2012. Proceedings

Victor S. Frost and Benjamin Melamed. "Traffic Modelingfor Telecommunications Networks." IEEE Communications, Mar. 1994.

Vinary J. Ribeiro, Rudolf H. Riedi, Matthew S.Crouse, and Richard G. Baraniuk. "Simulation of non-Gausian Long-Rang-Dependent Traffic Using Wavelets."In Proceeding SIGMETRICS '99, April 9,1999.

Watson H.W. "Solution to Problem 4001." Educational Times, August 1, 1873.